ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computer Speech and Language 17 (2003) 137–152

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

A probabilistic framework for segment-based speech recognition

James R. Glass *

MIT Laboratory for Computer Science, 200 Technology Square, Cambridge, MA 02139, USA

Received 10 December 2001; received in revised form 1 August 2002; accepted 3 November 2002

Abstract

Most current speech recognizers use an observation space based on a temporal sequence of measurements extracted from fixed-length "frames" (e.g., Mel-cepstra). Given a hypothetical word or sub-word sequence, the acoustic likelihood computation always involves *all* observation frames, though the mapping between individual frames and internal recognizer states will depend on the hypothesized segmentation. There is another type of recognizer whose observation space is better represented as a network, or graph, where each arc in the graph corresponds to a hypothesized variable-length segment that is represented by a fixed-dimensional "feature". In such feature-based recognizers, each hypothesized segmentation will correspond to a segment sequence, or path, through the overall segment-graph that is associated with a *subset* of all possible feature vectors in the total observation space. In this work we examine a maximum a posteriori decoding strategy for feature-based recognizers and develop a normalization criterion useful for a segment-based Viterbi or A^* search. Experiments are reported for both phonetic and word recognition tasks.

1. Introduction

Over the past two decades, first-order hidden Markov models (HMMs) have emerged as the dominant stochastic model for automatic speech recognition (ASR) (Rabiner, 1989). With a well-formed mathematical foundation, and efficient, automated training procedures which can process the ever increasing amounts of speech data, impressive HMM-based recognizers have been created for a wide-variety of increasingly difficult ASR tasks.

* Tel.: +1-617-253-1640; fax: +1-617-258-8642.
E-mail address: glass@mit.edu.

0885-2308/03/$ - see front matter
doi:10.1016/S0885-2308(03)00006-8

While HMMs have shown themselves to be highly effective, it is reasonable to question some of their basic structure. For example, almost all HMM-based ASR formulations restrict their acoustic modelling to an observation space defined by a temporal sequence of feature vectors computed at a fixed frame rate, typically once every 10 ms. As a result, adjacent feature vectors, especially those within the same phonetic segment, often exhibit smooth dynamics and are highly correlated, violating the conditional independence assumption imposed by the HMM model (Digilakis, 1992). The relationship between features computed in different phonetic segments is weaker, however. These observations motivate a framework which makes fewer conditional independence assumptions between observation frames; especially for those occurring within a phonetic segment.

A second property of conventional HMM-based formulations is their use of homogeneous, frame-based feature vectors such as Mel-frequency cepstral coefficients (MFCCs). Such a representation may not be able to adequately capture certain acoustic measurements known to be important for phonetic distinctions. This narrow view of acoustic-modelling can be quite constraining, and can make it difficult to incorporate acoustic-phonetic information into the decoding process. For example, it was with heterogeneous information sources and classifiers that we were able to achieve state-of-the-art phonetic *classification* results on the TIMIT corpus (Halberstadt and Glass, 1998), not with a frame-based processing technique.

Many speech scientists believe that the acoustic cues important for phonetic contrasts are best characterized in relation to specific temporal landmarks in the speech signal, such as points of oral closure (or release), or other points of maximal constriction (or opening) in the vocal tract which are produced during speech production (Stevens, 1995). Many of these locations correspond to phonetic boundaries, leading some speech researchers to consider segment- and landmark-based approaches for ASR.

In the past, there have been many segment-based ASR approaches which extracted feature vectors at specific temporal landmarks (Cole et al., 1983), including work during the early ARPA-SUR project in the 1970s (Weinstein et al., 1975). Most of these efforts were hampered however, by attempting to explicitly incorporate speech knowledge by heuristic means through intense knowledge engineering, and by lack of a stochastic framework to deal with the present state of ignorance in our understanding of the human communication process and its inherent variabilities. For this reason, much of the more recent work on segment-based speech recognition has largely avoided these issues, and has used decoding strategies similar to HMM-based approaches, including modelling acoustic likelihoods on a frame-by-frame basis (Ostendorf et al., 1996). For the purposes of this paper, these approaches are also considered to be 'frame-based'.

Although segment-based and landmark-based approaches have been largely eclipsed by the more powerful frame-based HMM paradigm and associated learning machinery, it is reasonable to ask: Is it the HMM structure which is so powerful, or the well-formulated training and decoding algorithms? If segment-based recognizers also made use of a probabilistic framework and training algorithms, how would their performance compare to HMM capabilities?

The SUMMIT speech recognizer developed by our group has always used a segment-based framework for its acoustic-phonetic representation of the speech signal (Zue et al., 1989; Glass et al., 1996). As illustrated in Fig. 1, acoustic or probabilistic landmarks form the basis for a phonetic segment network, or graph. Feature vectors are extracted both over hypothesized phonetic segments and at their boundaries for phonetic analysis. The resulting observation space

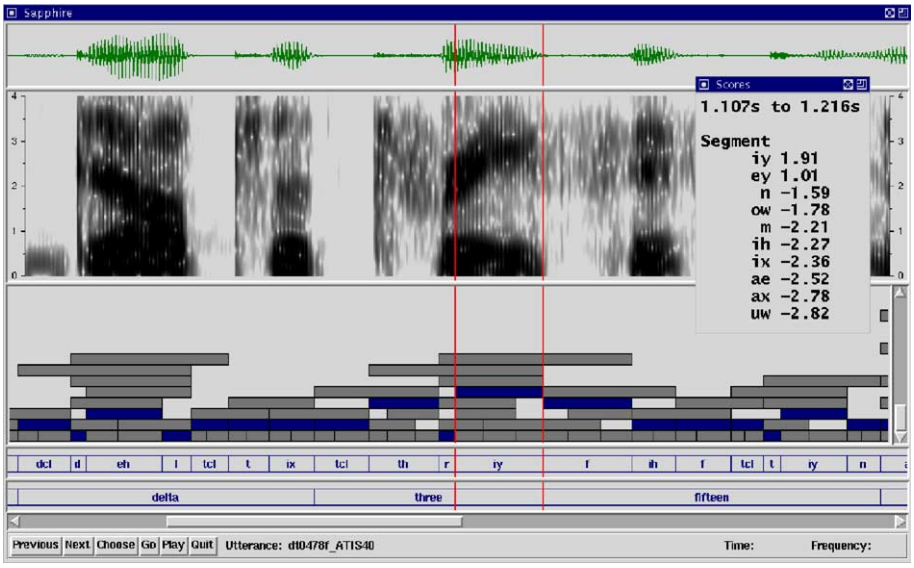

Fig. 1. Graphical displays from the SUMMIT segment-based speech recognizer. The top two displays contain the speech waveform and associated spectrogram, respectively. Below them, a segment-network display shows hypothesized phonetic segments; each segment spans a time range. The darker colored segments show the segmentation which achieved the highest score during search. The two transcriptions below the segment-network contain the best-scoring phonetic, and word sequences, respectively. The pop-up menu shows the ranked log-likelihood ratio segment scores for the [i] in the word "three".

(the set of all feature vectors) takes the form of an acoustic-phonetic network, or graph, whereby different paths through the graph are associated with different sets of feature vectors. This graph-based observation space is quite different from prevailing approaches which employ a temporal *sequence* of observations, which typically contain short-time spectral information (e.g., MFCCs). The segmental and feature-extraction characteristics of this recognizer provide us with a framework within which we try to incorporate knowledge of the speech signal. They enable us to explore different strategies for extracting information from the speech signal, and allow us to consider a much larger variety of observations than we could with traditional frame-based observations.

As illustrated in Fig. 2, the conventional MAP decoding framework must be modified for a graph-based observation space, since each path through the graph is associated with a different set of acoustic observations. In a conventional frame-based ASR decoder, acoustic likelihoods for any hypothesized word and sub-word segmentation are generated from the same underlying observation sequence of spectral frames. In a graph-based ASR decoder, different paths through the graph compute likelihoods on different observation spaces, because each arc has a unique acoustic observation. Thus, it is not possible to directly compare the likelihoods computed from different paths to decide on the most-likely word sequence. A normalization criterion is required to compare different paths.

Another interesting property of graph-based observation spaces is that some of the observations will correspond to segments which have no valid interpretation as lexical units (i.e., too long, too short, etc.). In Fig. 2, these observations correspond to the white (non-shaded) segments, and

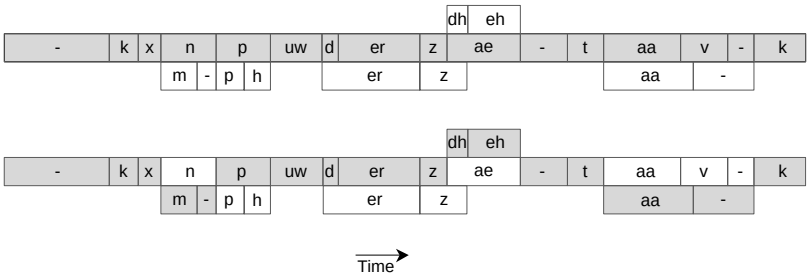

Fig. 2. A hypothetical segment-based graph for an utterance "computers that talk" illustrating two (shaded in gray) of 48 possible segmentations through the graph. Each segment in the graph, which corresponds to a hypothesized phonetic unit with a given start and end time, is associated with a single acoustic observation. Since the total observation space is the collection of all individual acoustic observations, each segmentation (a contiguous sequence of segments which span the entire utterance) is therefore associated with a different subset of all the observations in the graph.

can be considered to be *negative* examples of phones. Since these non-phonetic segments will be considered as possible phones during decoding, it is important to model them in addition to modelling the positive examples of conventional phones (e.g., the shaded segments in the lower path). If no negative phone model(s) are available, non-phonetic segments could only be scored as a conventional phone during decoding; an inconsistency that would not have a meaningful interpretation and could have negative consequences on the overall path score. This problem is especially serious if likelihoods are converted to posterior probabilities, since a poor likelihood could result in a very good posterior probability only because it happens to be a little better than the alternatives. As will be shown in the next section, proper modelling of negative examples will provide the key to a useful score normalization of graph-based observation spaces.

In this paper we describe the graph-based decoding mechanisms used in our stochastic segment-based ASR research. The recognizer utilizes the entire network of hypothesized segments (both positive and negative examples) during training, and accounts for the entire observation space during decoding. In the next section we describe the standard MAP decoding framework, and then derive three different mechanisms we have used to implement efficient decoding for our segment-based recognizer. We then report experimental evidence on phonetic and word recognition that we have used to evaluate the framework.

2. MAP decoding

In most probabilistic formulations of speech recognition the goal is to find the sequence of words $W^* = w_1, \ldots, w_N$, which has the maximum a posteriori (MAP) probability $P(W|A)$, where A is the set of acoustic observations associated with the speech waveform:

$$W^* = \arg\max_W P(W|A). \tag{1}$$

In most (large vocabulary) speech recognizers, decoding is accomplished by hypothesizing (usually implicitly) a segmentation, S, of the waveform into a sequence of sub-word states or units, U. Eq. (1) can then be rewritten as:

$$W^* = \arg\max_W \sum_{\forall S,U} P(S,U,W|A) \approx \arg\max_{S,U,W} P(S,U,W|A). \quad (2)$$

The latter approximation assumes that there is a single "correct" segmentation, S^*, associated with unit sequence, U^*, and words W^*. The approximation simplifies decoding by allowing the use of dynamic programming or graph search algorithms which seek only the "best" path (e.g., Viterbi, or A^*).

The expression for $P(S,U,W|A)$ is typically converted to the form:

$$P(S,U,W|A) = \frac{P(A|S,U,W)P(S|U,W)P(U|W)P(W)}{P(A)}. \quad (3)$$

Since $P(A)$ is independent of S, U, and W, it will not affect the outcome of the search, and is usually ignored. The term $P(U|W)$ can be considered to be a pronunciation model which predicts the probability of a sequence of sub-word units, U, being generated by the given word sequence, W. In most recognizers this is accomplished via a dictionary lookup. Some speech recognizers do incorporate a stochastic component at this level to attempt to model phonological variations in the way a word can be realized in fluent speech (e.g., "did you" being realized as the phoneme sequence /dɪǰu/) (Zue et al., 1989; Riley and Ljolje, 1991; Hetherington, 2001). The term $P(S|U,W)$ models the probability of the segmentation itself, and typically depends only on U. In HMMs for example, this term corresponds to the likelihood of a particular state sequence, and is generated by the state transition probabilities (which are ignored in some HMM decoders). More generally however, this term can be considered as a duration model which predicts the probability of individual segment durations. Many researchers have explored the use of more explicit duration models for recognition, especially in the context of segment-based recognition (Ostendorf and Roucos, 1989; Livescu and Glass, 2001). The remaining two terms, $P(A|S,U,W)$ and $P(W)$, correspond to the well-known acoustic and language models, respectively, with the former being the focus of this paper. Assuming A is conditionally independent of W given U, $P(A|S,U,W) = P(A|S,U)$. The following sections describe the acoustic modelling issues for frame-based and graph-based observations in more detail.

2.1. Frame-based observations

As described earlier, most ASR decoders take as input a temporal sequence of spectral vectors or frames, $O = \{o_1, \ldots, o_T\}$, which are normally computed at regular time intervals (e.g., every 10 ms). When the observation space consists of this sequence of frames, $A = O$, and acoustic likelihoods are computed for *every* frame during decoding. Thus, the term $P(A|S,U)$ accounts for *all* observations, and competing word hypotheses can be compared directly to each other since their acoustic likelihood is derived from the same observation space.

All recognizers that use frame-based observations as input to the decoder fit into this framework, including all discrete and continuous observation HMMs, and those using artificial neural networks for classification (Lamel & Gauvain, 1993; Mari, Fohr, & Junqua, 1996; Robinson, 1994; Young & Woodland, 1994; Robinson, Hochberg, & Renals, 1994). Many segment-based techniques also use a common set of sequential observations as well. Marcus for example, predetermined a set of acoustic-phonetic sub-segments, represented each by an observation vector,

which was then modelled with an HMM (Marcus, 1993). Other segment-based techniques hypothesize segments, but compute likelihoods on frames (Roucos, Ostendorf, Gish, & Derr, 1988; Digilakis, Rohlicek, & Ostendorf, 1993; Russell, 1993; Ljolje, 1994; Holmes & Russell, 1996).

2.2. Graph-based observations

In contrast to frame-based approaches, in a *graph*-based framework, each variable-length segment, s_i, is represented by a single fixed-dimensional feature vector, x_i. The observation space, A, consists of all the feature vectors in a segment network, and is thus significantly different from a frame-based ASR framework. Typically, there is an extra stage of processing to convert the frame sequence O to corresponding segmental feature vectors. Explicit segment or boundary hypotheses are typically used to compute the observation. A given n unit segmentation $S = \{s_1, \ldots, s_n\}$ will have a set of corresponding n observations $X = \{x_1, \ldots, x_n\}$. Figs. 2 and 3 contain hypothetical segment-based graphs, and illustrate two different segmentations, or paths, through each graph. Each segmentation, S, forms a contiguous sequence of segments spanning the entire graph, and is associated with a different subset of observations, X, in the graph.

Since alternative segmentations will consist of *different* observation sub-spaces, it is incorrect to compare the resulting likelihoods, $P(X|S,U)$, directly. In order to compare two paths we must consider the *entire* observation space. Thus, in addition to the observations X associated with the segmentation S, we must consider all other possible observations in the space Y, corresponding to the set of all other possible segments, R. For any given segmentation, S, X, and Y are mutually exclusive and collectively exhaustive, so that $X \cap Y = \emptyset$, and $X \cup Y = A$. In the top path in Fig. 3 for example, $X = \{a_1, a_3, a_5\}$, and $Y = \{a_2, a_4\}$. In the bottom path, $X = \{a_1, a_2, a_4, a_5\}$, and $Y = \{a_3\}$. The total observation space A, contains *both* X and Y, so for MAP decoding it is necessary to estimate $P(X, Y|S,U)$. Note that since X implies S, we can write $P(X, Y|S,U) = P(X, Y|U)$. The following two sections discuss methods for estimating the latter term in an efficient manner.

2.3. Modelling non-lexical units with an anti-phone

One approach to modelling $P(X, Y|U)$ is to add an extra lexical unit that is defined to map to all segments which do *not* correspond to one of the existing units in U. Consider the case where

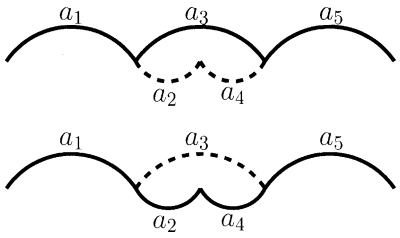

Fig. 3. Two segmentations (in solid lines) through a simple five segment graph with acoustic observations $\{a_1, \ldots, a_5\}$. The top segmentation is associated with observations $\{a_1, a_3, a_5\}$, while the bottom segmentation is associated with $\{a_1, a_2, a_4, a_5\}$.

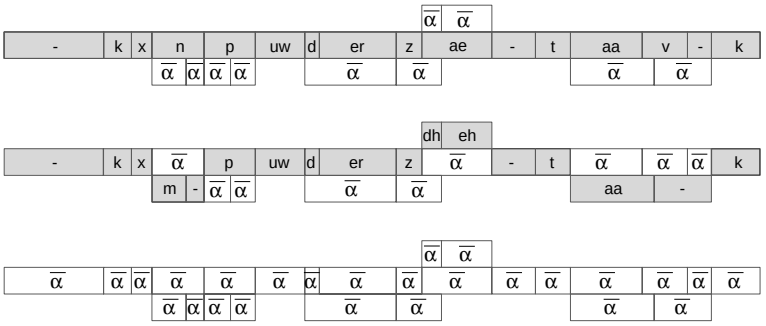

Fig. 4. The anti-phone normalization method assigns all segments not in a hypothesized segmentation to the anti-phone unit, $\bar{\alpha}$. The top two graphs show the anti-phone assignments made for the segmentations in Fig. 2. The bottom graph illustrates the concept of assigning every segment to be the anti-phone. Since this likelihood is a constant for any given graph, computation can be reduced to a simple likelihood ratio (Eqs. 4 and 5).

acoustic-modelling is done at the phonetic level, so that we build probabilistic models for individual phones, $\{\alpha\}$. In this approach we can view the segments in R as corresponding to the extra *anti-phone* unit, $\bar{\alpha}$. This unit represents all types of sounds which are *not* a phone as they are either too large, too small, or overlapping, etc. As shown in Fig. 4, two competing paths must therefore account for *all* segments, either as normal sub-word units in U or as the anti-phone $\bar{\alpha}$.

As illustrated in Fig. 4, we can avoid classifying all the segments in the search space by recognizing that $P(X, Y|\bar{\alpha})$, the probability that *all* segments are not a lexical unit, is a constant for any given graph, and therefore has no effect on decoding. If we also make the more dubious assumption of conditional independence between X and Y, given U, and note that $P(Y|U)$ depends only on $\bar{\alpha}$, we can decompose and rearrange $P(X, Y|U)$ as follows:

$$P(X, Y|U) = P(X|U)P(Y|\bar{\alpha})\frac{P(X|\bar{\alpha})}{P(X|\bar{\alpha})} \propto \frac{P(X|U)}{P(X|\bar{\alpha})}. \tag{4}$$

Thus, when we consider a particular segmentation S we need only concern ourselves with the n observations corresponding to S, but we must combine *two* terms for each segment s_i. The first term is the standard phonetic likelihood $P(x_i|U)$ (or more typically, $P(x_i|u_i)$). The second term is the likelihood that the segment is the anti-phone unit, $P(x_i|\bar{\alpha})$. The net result which must be maximized during search is:

$$W^* = \arg\max_{S,U,W} \prod_{i=1}^{n} \frac{P(x_i|u_i)}{P(x_i|\bar{\alpha})} P(s_i|u_i)P(U|W)P(W). \tag{5}$$

Eq. (5) assumes conditional independence between individual segmental feature vectors, x_i, given a phonetic hypothesis, U. While it is likely to be more reasonable to assume conditional independence between phonetic segments than between individual frames, there are many phonetic sequences (e.g., vowels/semi-vowels) that are highly correlated, and would be better modelled jointly. Note however, that the nature of the linguistic unit, u_i, could be designed to take common correlations into account. For example, we have successfully modelled highly correlated phonetic sequences such as /ar/ and /ir/, using this approach.

The conditional independence assumption made in Eq. (4) between the observations in X and Y, given U is also suspect. Segments that temporally overlap with each other will surely have correlated feature vectors to some degree. A better assumption might be to partition the segment network into sequential sub-graphs that could more reasonably be considered conditionally independently from each other. One such approach, called near-miss modelling, is described in the following section.

2.4. Beyond anti-phones: near-miss modelling

Anti-phone modelling partitions the observation space into essentially two parts: segments which are in a hypothesized segmentation, S, and those in R which are not. The anti-phone model is quite general since it must model all examples of observations which are not valid units. A larger, context-dependent inventory of anti-phone models might better model observations which are near-misses of particular units. One such method, called *near-miss* modelling, was developed by Chang (Chang and Glass, 1997; Chang, 1998). Near-miss modelling works by dividing the feature vector observation space, A, into many different subsets, such that there is one subset, A_i, associated with each segment, s_i. [1] The feature vectors (and associated segments) in A_i can therefore be viewed as near-misses of the feature vector x_i (for segment s_i). The top graph in Fig. 5 illustrates two possible near-miss subsets for two segments in a segment graph.

If near-miss subsets are carefully chosen, so that the subsets along any particular segmentation, S, are mutually exclusive, and collectively exhaustive (i.e., $A_i \cap A_j = \emptyset \quad \forall s_i, s_j \in S, i \neq j$, and $A = \bigcup(x_i \cup A_i)$), then an effective decoding strategy can be employed. During recognition, all observations in X along the hypothesized segmentation S are classified as normal via $P(X|U)$. The observations in Y are partitioned into the corresponding near-miss subsets, A_i, of the segments in S. The net result can be represented as:

$$W^* = \arg\max_{S,U,W} \prod_{i=1}^{n} P(x_i|u_i)P(A_i|u_i)P(s_i|u_i)P(U|W)P(W), \tag{6}$$

where $P(A_i|u_i)$ is computed as the product of observations in A_i being generated by the near-miss model associated with u_i (i.e., $\bar{u}_i$). In the simplest case the near-miss model might just be the anti-phone model. However, it could also depend on the sub-word unit being hypothesized for that segment. The bottom graph of Fig. 5 illustrates the near-miss subsets associated with one particular segmentation of a segment graph.

One of Chang's key insights was that a simple temporal criterion could be used to determine near-miss subsets. This is because any particular segmentation through a segment network accounts for all times exactly once. Thus, segments that span a *common point in time* are natural near-misses of each other, since only *one* of them can be a member of a particular segmentation, S. Using a common reference point, such as the segment mid-point, appropriate near-miss subsets can be defined which satisfy the necessary near-miss conditions. This near-miss subset criterion is illustrated in the top graph of Fig. 5.

[1] Note that A_i could be an empty set.

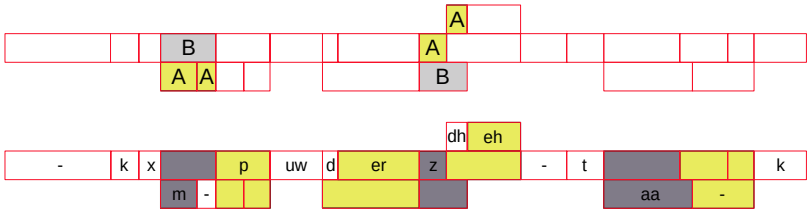

Fig. 5. Near-miss subsets are computed for every segment in the segment graph. A temporal criterion can be used to effectively determine near-miss subsets. In the top graph, the segments labelled A are considered to define the near-miss subsets of segments labelled B because their mid-point is spanned by B (other criterion could also be used). If, as part of some segmentation, segment B was considered to be sub-word unit α, the associated A segments would be scored by the near-miss model for α. The lower graph illustrates the near-miss partitioning occurring for one particular segmentation of the segment graph. Each segment with a phonetic label is part of the hypothesized segmentation; its associated near-miss subset (if it exists) is shaded the same amount.

The near-miss modelling method makes similar conditional independence assumptions between segments as was made by the anti-phone method. However, the near-miss modelling framework has the potential for more sophisticated modelling of a near-miss subset likelihood, $P(A_i|u_i)$, than could be achieved by the anti-phone method. Furthermore, a conditional independence assumption between contiguous near-miss subsets would seem to be more reasonable than conditional independence between segments (which in turn is more reasonable than conditional independence between frames).

2.5. Modelling landmarks

In addition to modelling segments, it is often desirable to provide additional information about segment boundaries, or landmarks (Phillips and Glass, 1994). If we call the acoustic observations extracted at landmarks, Z, we must now consider the joint space $A = XYZ$ as our observation space. It thus becomes necessary to estimate the probability $P(X, Y, Z|S, U)$. If we assume conditional independence between the observations XY representing segments, and Z representing landmarks, we can write:

$$P(X, Y, Z|S, U) = P(X, Y|S, U)P(Z|S, U). \tag{7}$$

If Z corresponds to a set of observations taken at landmarks or boundaries, then a particular segmentation will assign some of the landmarks to *transitions* between lexical units, while the remainder will be considered to occur *internal* to a unit (i.e., within the boundaries of a hypothesized segment). This concept is illustrated in Fig. 6. Since any segmentation accounts for *all* of the landmark observations Z, there is no need for a normalization criterion such as was discussed for graph-based observations, although it can be useful to normalize likelihoods by $P(Z)$ (such as for confidence scoring (Hazen et al., 2002)). If we assume conditional independence between the m individual observations in Z given U, $P(Z|S, U)$ can be written as

$$P(Z|S, U) = \prod_{i=1}^{m} P(z_i|S, U), \tag{8}$$

where z_i is the observation extracted at the ith landmark.

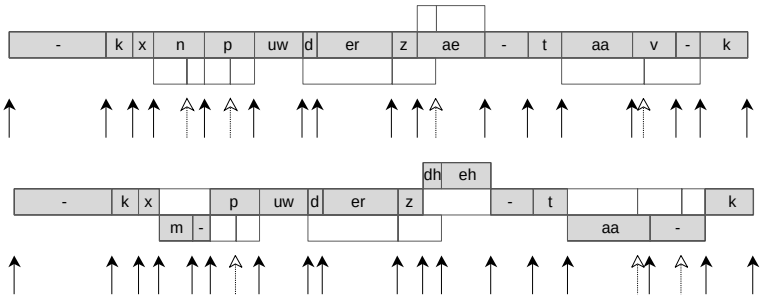

Fig. 6. As a complement to graph-based observations, acoustic modelling can also be performed at hypothesized segment boundaries, or landmarks. This figure shows two segmentations through the same segment graph. The solid arrows point to hypothesized phone *transitions*, while the hollow arrows point to landmarks which are phone-*internal*. Note that the assignments are segmentation dependent, and that all landmarks are accounted for in any segmentation through the graph.

The landmark modelling framework makes similar kinds of assumptions as the anti-phone and near-miss methods in that it assumes conditional independence between individual landmarks, and between all segment- and landmark-based measurements, XY and Z. Landmarks are comparable to frames in that they are sequential in nature, although they are non-uniformly distributed in time, and occur much less frequently. Thus, the conditional independence assumption between individual landmarks is arguably more reasonable than that made in frame-based methods. The case for conditional independence between landmark and segmental measurements is weaker, although the measurements are typically different; segmental measurements tend to quantify intra-segment properties (e.g., within-phone dynamics) whereas landmark measurements tend to quantify inter-segment properties (e.g., transitions between phones).

2.6. Decoding implementation

Recognition with our segment-based framework is done via a modified Viterbi algorithm that can be viewed as finding the best path through two graphs; a conventional pronunciation graph representing all possible word sequences and their associated pronunciations, and an acoustic-phonetic graph representing all possible segmentations of a spoken utterance. Our current recognizer is implemented in terms of weighted finite-state transducers (Glass et al., 1999). From this perspective, we view recognition as finding the best path through the composition $A \circ U$, where A and U represent the acoustic-phonetic, and pronunciation graphs, respectively. The pronunciation graph itself represents the composition $U = C \circ P \circ L \circ G$, where C converts context-dependent to context-independent lexical labels, P applies phonological rules, L maps pronunciations to words in the lexicon, and G is the language model.

Although one could explore an exhaustive segmentation method (i.e., one that could consider all possible segmentations), we have typically used explicit segmentation algorithms in our research, mainly because of our interest in real-time ASR for conversational interfaces (Zue et al., 2000). Our original work in this area used a hierarchical acoustic clustering method for determining important landmarks and associated regions in the speech signal (Glass and Zue, 1988; Glass, 1988). More recently we have used ASR methods to produce a probabilistic phonetic segmentation (Chang and Glass, 1997; Lee and Glass, 1998).

Our recognizer uses a Viterbi-style training, whereby forced alignments of orthographic transcriptions are computed to create reference phonetic transcriptions that are used to train acoustic-phonetic models. Context-dependent landmark models are trained based on the forced alignment segmentations. Inter-segmental landmarks are modelled as transitions between two phones, whereas intra-segmental landmarks are modelled as internal boundaries for the given phone. Anti-phone and near-miss models are trained on all segments that are *not* part of a forced alignment segmentation through a segment graph (i.e., they are negative examples of phones). Anti-phone models can be trained on all negative examples. For near-miss modelling, the negative examples are further divided into near-miss subsets using the temporal criterion described in Section 2.4. Each of the labelled subsets are used to train the appropriate near-miss model.

3. Experiments

Although the main intent of this paper is to describe the probabilistic framework we have developed for our segment-based recognizer, we will also present some experimental results which quantify its performance. Over the years we have evaluated our recognizer on many different tasks (Chang, 1998; Livescu and Glass, 2001), but will present two experiments we have performed on phonetic and word recognition (Halberstadt and Glass, 1998; Ström et al., 1999).

3.1. Recognizer configuration

All experiments have made use of the SUMMIT segment-based speech recognition system which combines segment- and landmark-based classifiers. Feature extraction is typically based on averages and derivatives of MFCCs, plus additional information such as energy. Acoustic models are based on mixtures of diagonal Gaussians. Prior to acoustic modelling, our feature space is whitened with a global rotation which transforms the pooled within-class covariance of the training data to an identity matrix. The rotation vectors are composed of: (1) the inverse standard deviations of the original feature space, (2) the eigenvectors of the corresponding pooled within-class correlation coefficients, and (3) the inverse square roots of the corresponding eigenvalues. With this method the pooled within-class distribution of the normalized training data will be uncorrelated, and have unity variance across all dimensions. While the statistical properties of individual phonetic classes can have strong correlations, we have observed small, but consistent improvements in recognition performance with this transformation. This technique can also be used to reduce the dimensionality of the feature vector if desired.

3.2. Phonetic recognition

Our phonetic recognition experiments were based on the widely used TIMIT acoustic-phonetic corpus (Garofolo et al., 1990). Acoustic and language models were built using the provided 61 label set. Recognition errors were tabulated using the 39 labels used by others to report recognition results (Lee and Hon, 1989; Lamel and Gauvain, 1993; Robinson, 1994). Models were trained on the designated training set of 462 speakers, and results are reported on the 24 speaker core test set. A 50 speaker development set (taken from the remaining 144 speakers in the full test

Table 1
Segment-based TIMIT phonetic recognition error rates (Halberstadt and Glass, 1998)

Method	Phone error rate (%)
Segment-based, Anti-phone	27.7
Segment-based, Near-miss	26.4
Landmark-based	24.9
Landmark + Near-miss	24.8
Landmark + Anti-phone	24.4

Table 2
Reported phonetic recognition error rates on the TIMIT core test set

Method	Phone error rate (%)
Triphone CDHMM (Lamel and Gauvain, 1993)	27.1
Recurrent Neural Network (Robinson, 1994)	26.1
Bayesian Triphone HMM (Ming and Smith, 1998)	25.6
Near-miss, probabilistic segmentation (Chang, 1998)	25.5
Anti-Phone, Heterogeneous classifiers (Halberstadt and Glass, 1998)	24.4

set) was used for intermediate experiments so that the core test set was used only for final testing. Reported results are phonetic recognition error which includes substitution, deletion, and insertion errors. The language model used in all experiments was a phone bigram based on the training data with perplexity 15.8 on the development set (using 61 labels). A single parameter (optimized on the development set) controlled the trade-off between insertions and deletions.

The segmental and landmark representations used in these experiments were based on five variations of averages and derivatives of 12 Mel-frequency and PLP cepstral coefficients, plus energy and duration (Halberstadt, 1998; Halberstadt and Glass, 1998). Fourfold aggregation was used to improve the robustness of the Gaussian mixtures (Hazen and Halberstadt, 1998). A probabilistic segmentation was used to produce segment graphs with a density of approximately 60 segments/s, compared to the 13 segments/s found in an average TIMIT phonetic transcription (Chang, 1998). The results of different experiments with a committee-based classifier to combine the outputs of five different segmental, and landmark models are shown in Table 1. Table 2 compares the best result with the best results reported in the literature.

The results in Table 1 indicate that when only context-independent segment models were used for recognition, the near-miss model outperforms the anti-phone model. The table also shows that context-dependent landmark models do better than either context-independent method. This is not surprising result since there are significantly more acoustic models. However, despite the difference in size, the segmental models do improve the overall performance when they are used in tandem with the landmark models. We have not yet attempted context-dependent segment models for the task of phonetic recognition. However, the next section reports the results of combining context-dependent segment and landmark models for word recognition.

3.3. Word recognition

Word recognition experiments have been performed on a spontaneous-speech, telephone-based, conversational interface task in the weather domain (Glass et al., 1999). For these ex-

Table 3
Word recognition error rates for weather domain task

Method	Word error rate (%)
Segment models	9.6
Landmark models	7.6
Combined	6.1

periments, a 50 000 utterance training set was used, and an 1806 utterance test set was used containing in-domain queries. The recognizer for this task used a vocabulary of 1957 words, as well as a class bigram and trigram language model with perplexities of 17.6 and 15.9, respectively (Ström et al., 1999). As shown in Table 3, the word error rate (WER) obtained by the system using only context-dependent segmental models was 9.6%. When landmark models were used, the WER decreased to 7.6%. The overall performance decreased to 6.1% when both models were combined.

4. Discussion

The results shown in Table 2 compare a number of published results on phonetic recognition which have been based on the TIMIT core test set. There are still differences regarding the complexity of the acoustic and language models, thus making a direct comparison somewhat difficult. Nevertheless, we believe our results are competitive with those obtained by others, which makes a strong case for the viability of a segment-based approach.

One of the nice properties of our recognizer is its fast training cycle. By distributing computation among approximately 10–15 Pentium-based processors, we can perform one complete acoustic training iteration on 100 h of speech data in about 5–6 h. Since our recognizer converges with very few iterations we can therefore completely retrain our recognizer on a large speech corpus in less than a day. This compares very favorably to many other HMM-based systems which can take several weeks to completely retrain. Part of the reason for our training speed is that we use Viterbi-style training rather than the more conventional Baum–Welch training. While we believe the Viterbi approximation is reasonable for a segment-based framework (since different paths must differ by a complete segment rather than by just a frame), we plan to explore training methods which consider more alternatives than a single path (e.g., a constrained form of Baum–Welch training).

The framework we have outlined in this paper provides flexibility to explore the relative advantages of segment versus landmark representations. As we have shown, it is possible to use only segment-based feature vectors, or landmark-based feature vectors (which could reduce to frame-based processing), or a combination of both. Since the acoustic modelling methods described in this paper are quite different from conventional frame-based approaches, it is quite possible that the two approaches contain complementary information. Thus it might be possible that a technique which combined the two methods, either at the sub-word, word, or sentence levels, could gain additional improvement.

The anti-phone normalization criterion can be interpreted as a likelihood ratio. In this way it has similarities with techniques being used in word-spotting, which compare acoustic likelihoods with those of "filler" models (Rohlicek et al., 1989; Wilpon et al., 1990). The likelihood or odds

ratio was also used by Cohen to use HMMs for segmenting speech (Cohen, 1981). The landmark models used in this framework have similarities with variable-frame-rate analysis (Zhu and Alwan, 2000), though our landmark feature vectors are typically much longer in length (e.g., spanning 150 ms), and occur less frequently.

Although the graph-based decoding framework was developed for segment-based speech recognition, it could apply to any pattern recognition problem that used a graph-based observation space. One of the biggest assumptions made in this work was the conditional independence assumption made between the observations in X, and those in Y. Segments that temporally overlap with each other are clearly related to each other to some degree. In the future, it would be worthwhile examining alternative methods for modelling this joint XY space.

5. Summary

In this paper we have described a probabilistic decoding framework for decoding a graph-based observation space. This method is particularly appropriate for segment-based speech recognizers which transform the observation space from a sequence of frames, to a graph of features. Graph-based observation spaces allow for a wider variety of modelling methods to be explored than could be achieved with frame-based approaches. We have developed two methods for decoding graph-based observation spaces, based on anti-phone or near-miss modelling, and have achieved good results on phonetic and word recognition tasks.

Acknowledgements

There are a number of colleagues, past and present, who have contributed to this work including Jane Chang, Andrew Halberstadt, T.J. Hazen, Lee Hetherington, Michael McCandless, Nikko Ström, and Victor Zue. This paper was considerably improved by the thoughtful feedback from two anonymous reviewers. This research was supported by DARPA under contract N66001-99-C-1-8904 monitored through the Naval Command, Control and Ocean Surveillance Center.

References

Chang, J., 1998. Near-miss modeling: a segment-based approach to speech recognition. Ph.D. thesis, EECS, MIT.

Chang, J., Glass, J., 1997. Segmentation and modeling in segment-based recognition. In: Proc. Eurospeech, Rhodes, Greece, October, pp. 1199–1202.

Cohen, J., 1981. Segmenting speech using dynamic programming. J. Acoust. Soc. Am. 69 (5), 1430–1438.

Cole, R., Stern, R., Phillips, M., Brill, S., Pilant, A., Specker, P., 1983. Feature-based speaker-independent recognition of isolated letters. In: Proc. ICASSP, Boston, MA, pp. 731–733.

Digilakis, V., 1992. Segment-based stochastic models of spectral dynamics for continuous speech recognition. Ph.D. thesis, Boston University.

Digilakis, V., Rohlicek, J., Ostendorf, M., 1993. ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition. IEEE Trans. Speech Audio Proc. 1 (4), 431–442.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallet, D., Dahlgren, N., 1990. The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM. NTIS order number PB91-505065, October.

Glass, J., 1988. Finding acoustic regularities in speech: applications to phonetic recognition. Ph.D. thesis, EECS, MIT, May.
Glass, J., Chang, J., McCandless, M., 1996. A probabilistic framework for feature-based speech recognition. In: Proc. ICSLP Philadelphia, PA, pp. 2277–2280, October.
Glass, J., Hazen, T., Hetherington, L., 1999. Real-time telephone-based speech recognition in the Jupiter domain. In: Proc. ICASSP Phoenix, AZ, pp. 61–64, March.
Glass, J., Zue, V., 1988. Multi-level acoustic segmentation of continuous speech. In: Proc. ICASSP, New York, NY, pp. 429–432, April.
Halberstadt, A., 1998. Heterogeneous acoustic measurements and multiple classifiers for speech recognition. Ph.D. thesis, EECS, MIT, November.
Halberstadt, A., Glass, J., 1998. Heterogeneous measurements and multiple classifiers for speech recognition. In: Proc. ICSLP, Sydney, Australia, December, pp. 995–998.
Hazen, T., Halberstadt, A., 1998. Using aggregation to improve the performance of mixture Gaussian acoustic models. In: Proc. ICASSP, Seattle, WA, May, pp. 653–656.
Hazen, T., Seneff, S., Polifroni, J., 2002. Recognition confidence scoring and its use in speech understanding systems. Comp. Speech Lang. 16, 49–67.
Hetherington, L., 2001. An efficient implementation of phonological rules using finite-state transducers. In: Proc. Eurospeech, Aalborg, Denmark, September, pp. 1522–1609.
Holmes, W., Russell, M., 1996. Modeling speech variability with segmental HMMs. In: Proc. ICASSP, Atlanta, GA, May, pp. 447–450.
Lamel, L., Gauvain, J.L., 1993. High performance speaker-independent phone recognition using CDHMM. In: Proc. Eurospeech, Berlin, Germany, September, pp. 121–124.
Lee, K.F., Hon, H.W., 1989. Speaker-independent phone recognition using hidden Markov models. IEEE Trans. ASSP 37 (11), 1641–1648.
Lee, S., Glass, J., 1998. Real-time probabilistic segmentation for segment-based speech recognition. In: Proc. ICSLP, Sydney, Australia, December, pp. 1803–1806.
Livescu, K., Glass, J., 2001. Segment-based recognition on the PhoneBook task: initial results and observations on duration modeling. In: Proc. Eurospeech Aalborg, Denmark, September, pp. 1437–1440.
Ljolje, A., 1994. High accuracy phone recognition using context clustering and quasi-triphone models. Comput. Speech Lang. 8 (2), 129–151.
Marcus, J., 1993. Phonetic recognition in a segment-based HMM. In: Proc. ICASSP, Minneapolis, MN, April, pp. 479–482.
Mari, J.F., Fohr, D., Junqua, J.C., 1996. A second-order HMM for high performance word and phoneme-based continuous speech recognition. In: Proc. ICASSP, Atlanta, GA, May, pp. 435–438.
Ming, J., Smith., F., 1998. Improved phone recognition using Bayesian triphone models. In: Proc. ICASSP, Seattle, WA, May, pp. 409–412.
Ostendorf, M., Digilakis, V., Kimball, O., 1996. From HMM's to segment models: a unified view of stochastic modelling for speech recognition. IEEE Trans. Speech Audio Proc. 4 (5), 360–378.
Ostendorf, M., Roucos, S., 1989. A stochastic segment model for phoneme-based continuous speech recognition. IEEE Trans. ASSP 37 (12), 1857–1869.
Phillips, M., Glass, J., 1994. Phonetic transition modelling for continuous speech recognition. J. Acoust. Soc. Am. 95 (5), 2877.
Rabiner, L., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–286.
Riley, M., Ljolje, A., 1991. Lexical access with a statistically-derived phonetic network. In: Proc. Eurospeech Genoa, Italy, September, pp. 585–585.
Robinson, A., 1994. An application of recurrent nets to phone probability estimation. IEEE Trans. Neural Networks 5 (2), 298–305.
Robinson, T., Hochberg, M., Renals, S., 1994. IPA: improved phone modelling with recurrent neural networks. In: Proc. ICASSP, Adelaide, Australia, April, pp. 37–40.

Rohlicek, J., Russell, W., Roucos, S., Gish, H., 1989. Continuous hidden Markov modelling for speaker-independent word spotting. In: Proc. ICASSP, Glasgow, Scotland, May, pp. 627–630.

Roucos, S., Ostendorf, M., Gish, H., Derr, A., 1988. Stochastic segment modelling using the Estimate-Maximize algorithm. In: Proc. ICASSP, New York, NY, pp. 127–130.

Russell, M., 1993. A segmental HMM for speech pattern modelling. In: Proc. ICASSP, Minneapolis, MN, pp. 499–502.

Stevens, K., 1995. Applying phonetic knowledge to lexical access. In: Proc. Eurospeech, Madrid, Spain, pp. 3–11.

Ström, N., Hetherington, L., Hazen, T., Sandness, E., Glass, J., 1999. Acoustic modelling improvements in a segment-based speech recognizer. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop, Keystone, CO, December, pp. 139–142.

Weinstein, C., McCandless, S., Mondshein, L., Zue, V., 1975. A system for acoustic-phonetic analysis of continuous speech. IEEE Trans. ASSP 23, 54–67.

Wilpon, J., Rabiner, L., Lee, C.H., Goldman, E., 1990. Automatic recognition of keywords in unconstrained speech using hidden Markov models. IEEE Trans. ASSP 38 (11), 1870–1878.

Young, S., Woodland, P., 1994. State clustering in hidden Markov model-based continuous speech recognition. Comput. Speech Lang. 8 (4), 369–383.

Zhu, Q., Alwan, A., 2000. On the use of variable frame rate analysis in speech recognition. In: Proc. ICASSP, Istanbul, Turkey, June, pp. 1783–1786.

Zue, V., Glass, J., Phillips, M., Seneff, S., 1989. The MIT SUMMIT speech recognition system: a progress report. In: Proc. Speech and Natural Language Workshop, Philadelphia, PA, February, pp. 179–189.

Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., Hetherington, L., 2000. Jupiter: a telephone-based conversational interface for weather information. IEEE Trans. Speech Audio Proc. 8 (1), 85–96.